We claim:

1	1. A method for improving information retrieval, classification, indexing, and
2	summarization, comprising:
3	identifying a compound document as a coherent body of hyperlinked material on a single
4	topic as created by a number of collaborating authors;
5	analyzing the content and structure of the compound document to find a preferred entry point
<i>6</i>	for the compound document;
7	processing the compound document as a whole, including at least one of indexing,
8	classification, and retrieval; and
9	processing the compound document from the entry point, including at least one of creating at
10	least one of presentation of results from retrieval, summarization, and classification.
1	2. The method of claim 1 wherein the body of hyperliked material includes at least one of:
2	the internet, an intranet, and a digital library.
1	3. The method of claim 1 wherein the body of hyperlinked material is distributed over a
2	plurality of URLs.
1	4. The method of claim 1 wherein the identifying includes observing the results of a number
2	of heuristics run on the body of hyperlinked material and related hyperlinks.

3 6. The method of claim 4 wherein the heuristic includes identifying hyperlinks that contain

within the same directory and include a sufficient quantity of common anchor text.

5. The method of claim 4 wherein the heuristic includes identifying hyperlinks that link

- 4 linguistic structures that indicate relationships between parts of a document including at least
- one of a list of page numbers, and the terms "next", "previous", "index", "contents", and their
- 6 non-English equivalents.

1

2

- 7. The method of claim 4 wherein the heuristic includes identifying external hyperlinks to
- 2 the same places.
- 8. The method of claim 4 wherein the heuristic includes identifying at least one of: similar
- 2 creation dates and similar last-modified dates.
- 9. The method of claim 4 wherein the heuristic includes identifying individual URLs having
- 2 similar structure indicating an order of inclusion in the compound document.
- 10. The method of claim 4 wherein the heuristic includes identifying a link structure of
- 2 "wheel" form.
- 11. The method of claim 1 wherein the analyzing includes observing the results of a number
- of heuristics run on the component document and related hyperlinks.
- 1 12. The method of claim 11 wherein the heuristic includes identifying specific filenames
- that define the entry point, including at least one of: "index" and "default".
- 13. The method of claim 11 wherein the heuristic includes identifying a particular
- 2 component document in the compound document as the entry point because the component
- document has several in-links.
- 14. The method of claim 13 wherein the in-links are from outside the compound document.
- 15. The method of claim 11 wherein the heuristic includes identifying a particular
- 2 component document in the compound document as the entry point because the component
- document has several out-links.

1	16. The method of claim 11 wherein the heuristic includes determining a measure of vector
2	distances along intra-document links between a particular component document and all other
3	component documents in the compound document.
•	
1	17. The method of claim 11 wherein the heuristic includes determining whether a URL has
2	links pointing to longer URLs having common directory components followed by different
3	ending directory components.
1	18. The method of claim 17 wherein the ending directory components contain specific
2	identifying information.
1	19. The method of claim 11 wherein the analyzing includes combining the results of a
,	number of heuristics run on various component documents in the compound document,
2	wherein the results include numerical scores and the combining includes a weighted
<i>3</i>	averaging of the numerical scores into an overall score, and the maximum overall score
4	·
5	determines the preferred entry point.
1	20. A system for improving information retrieval, indexing, and summarization comprising:
2	a compound document identifier that detects a coherent body of hyperlinked material on a
3	single topic as created by a number of collaborating authors;
4	an analyzer that finds a preferred entry point for the compound document according to the
5	content and structure of the compound document; and
6	a compound document processor that performs at least one of indexing, classification, and
. 7	retrieval, for the compound document as a whole, and then performs at least one of
8	creating at least one presentation of results from retrieval, summarization, and

classification.

21. A computer program product instantiated on a computer-readable medium, comprising:
a first code means for identifying a compound document as a coherent body of hyperlinked
material on a single topic as created by a number of collaborating authors;
a second code means for analyzing the content and structure of the compound document to
find a preferred entry point for the compound document; and
a third code means for processing the compound document, wherein the processing includes
at least one of creating at least one presentation of results from retrieval,
aummerization, and alossification